# Limit theorems for sample eigenvalues in a generalized spiked population model

Zhidong Bai, Jian-Feng Yao

# LIMIT THEOREMS FOR SAMPLE EIGENVALUES IN A GENERALIZED SPIKED POPULATION MODEL

ZHIDONG BAI AND JIAN-FENG YAO

Abstract. In the spiked population model introduced by Johnstone [10], the population covariance matrix has all its eigenvalues equal to unit except for a few fixed eigenvalues (spikes). The question is to quantify the effect of the perturbation caused by the spike eigenvalues. Baik and Silverstein [6] establishes the almost sure limits of the extreme sample eigenvalues associated to the spike eigenvalues when the population and the sample sizes become large. In a recent work [5], we have provided the limiting distributions for these extreme sample eigenvalues. In this paper, we extend this theory to a *generalized* spiked population model where the base population covariance matrix is arbitrary, instead of the identity matrix as in Johnstone's case. New mathematical tools are introduced for establishing the almost sure convergence of the sample eigenvalues generated by the spikes.

## 1. Introduction

Let $(T_p)$ be a sequence of $p \times p$ non-random and nonnegative definite Hermitian matrices and let $(w_{ij})$, $i, j \geq 1$ be a doubly infinite array of i.i.d. complex-valued random variables satisfying

$$\mathbb{E}(w_{11}) = 0, \quad \mathbb{E}(|w_{11}|^2) = 1, \quad \mathbb{E}(|w_{11}|^4) < \infty.$$

1991 *Mathematics Subject Classification.* Primary 60F15, 60F05; secondary 15A52, 62H25.

*Key words and phrases.* Sample covariance matrices, Spiked population model, limit theorems, Largest eigenvalue, Extreme eigenvalues .

The research of this Zhidong Bai was supported by CNSF grant 10571020 and NUS grant R-155-000-061-112.

Research was (partially) completed while J.-F. Yao was visiting the Department of Statistics and Applied Probability, National University of Singapore in March 2007.

Write $Z_n = (w_{ij})_{1 \le i \le p, 1 \le j \le n}$, the upper-left $p \times n$ bloc, where $p = p(n)$ is related to $n$ such that when $n \to \infty$, $p/n \to y > 0$. Then the matrix $S_n = \frac{1}{n} T_p^{1/2} Z_n Z_n^* T_p^{1/2}$ can be considered as the sample covariance matrix of an i.i.d. sample $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of $p$-dimensional observation vectors $\mathbf{x}_j = T_p^{1/2} \mathbf{u}_j$ where $\mathbf{u}_j = (w_{ij})_{1 \le i \le p}$ denotes the $j$-th column of $Z_n$. Throughout the paper, $A^{1/2}$ stands for any Hermitian square root of an nonnegative definite (n.n.d.) Hermitian matrix $A$.

Assume that the empirical spectral distribution (ESD) of $T_p$ converges weakly to a nonrandom probability distribution $H$ on $[0, \infty)$. It is then well-known that the ESD of $S_n$ converges to a nonrandom limiting spectral distribution (LSD) $G$ [11, 13].

Let $\lambda_{n,1} \ge \cdots \ge \lambda_{n,p}$ be the set of sample eigenvalues, i.e. the eigenvalues of the sample covariance matrix $S_n$. The so-called *null case* corresponds to the situation $T_p \equiv I_p$, so that, assuming $y \le 1$, the LSD $G$ reduces to the Marčenko-Pastur law with support $\Gamma_G = [a_y, b_y]$ where $a_y = (1 - \sqrt{y})^2$ and $b_y = (1 + \sqrt{y})^2$. Furthermore, the extreme sample eigenvalues $\lambda_{n,1}$ and $\lambda_{n,p}$ almost surely tend to $b_y$ and $a_y$, respectively, and the sample eigenvalues $(\lambda_{n,j})$ fill completely the interval $[a_y, b_y]$. However, as pointed out by Johnstone [10], many empirical data sets demonstrate a significant deviation from this null case since some of sample extreme eigenvalues are well separated from an inner bulk interval. As a way for possible explanation of such phenomenon, Johnstone proposes a *spiked population model* where all eigenvalues of $T_p$ are unit except a fixed and relatively small number among them (*spikes*). In other words, the population eigenvalues $\{\beta_{n,j}\}$ of $T_p$ are

$$\underbrace{\alpha_1, \ldots, \alpha_1}_{n_1}, \ldots, \underbrace{\alpha_K, \ldots, \alpha_K}_{n_K}, \underbrace{1, \ldots, 1}_{p-M},$$

where $M$ is fixed as well as the multiplicity numbers $(n_k)$ which satisfy $n_1 + \cdots + n_K = M$. Clearly, this spiked population model can be viewed as a finite-rank perturbation of the null case.

Obviously, the LSD $G$ of $S_n$ is not affected by this small perturbation, still equals to the Marčenko-Pastur law. However, the asymptotic behavior of the extreme eigenvalues of $S_n$ is significantly different from the null case. The fluctuation of the largest eigenvalue $\lambda_{n,1}$ in case of complex Gaussian variables has been recently studied in Baik et al. [7]. These authors prove a transition phenomenon: the weak limit as well as the scaling of $\lambda_{n,1}$ is different according to its location with respect to a critical value $1+\sqrt{y}$. In Baik and Silverstein [6], the authors consider the spiked population model with general random variables: complex or real and not necessarily Gaussian. For the almost sure limits of the extreme sample eigenvalues, they also find that these limits depend on the critical values $1+\sqrt{y}$ for largest sample eigenvalues, and on $1-\sqrt{y}$ for smallest ones. For example, if there are $m$ eigenvalues in the population covariance matrix larger than $1+\sqrt{y}$, then the $m$ largest sample eigenvalues $\lambda_{n,1},\ldots,\lambda_{n,m}$ will converge to a limit above the right edge $b_y$ of the limiting Marčenko-Pastur law, see §4.1 for more details. In a recent work Bai and Yao [5], considering general random matrices as in [6], we have established central limit theorems for these extreme sample eigenvalues generated by spike eigenvalues which are outside the critical interval $[1-\sqrt{y}, 1+\sqrt{y}]$.

The spiked population model has also an extension to other random matrices ensembles through the general concept of small-rank perturbations. The goal is again to examine the effect caused on the sample extreme eigenvalues by such perturbations. In a series of recent papers [12, 9, 8], these authors establish several results in this vein for ensembles of form $M_n = W_n + n^{-1/2}V$ where $W_n$ is a standard Wigner matrix and $V$ a small-rank matrix.

The present work is motivated by a generalization of Johnstone's spike population model defined as follows. The population covariance matrix $T_p$ posses two sets of eigenvalues: a small number of them, say $(\alpha_k)$, called *generalized spikes*, are well separated - in a sense to be defined later-, from a base set $(\beta_{n,i})$. In other words, the spectrum of $T_p$ reads as

$$\underbrace{\alpha_1,\ldots,\alpha_1}_{n_1},\ldots,\underbrace{\alpha_K,\ldots,\alpha_K}_{n_K},\beta_{n,1},\ldots,\beta_{n,p-M}.$$

Therefore, this scheme can be viewed as a finite-rank perturbation of a general population covariance matrix with eigenvalues $\{\beta_{n,j}\}$.

The empirical distributions generated by the eigenvalues $(\beta_{n,i})$ will be assumed to have a limit distribution $H$. Note that $H$ is also the LSD of $T_p$ since the perturbation is of finite rank. Analogous to Johnstone's spiked population model, the LSD $G$ of the sample covariance matrix $S_n$ is still not affected by the spikes. The aim of this work is to identify the effect caused by the spikes $(\alpha_k)$ on a particular subset of sample eigenvalues. The results obtained here extend those of [6, 5] to the present generalized scheme.

The remaining sections of the paper are organized as following. §2 gives the precise definition of the generalized spiked population model. Next, we use §3 to recall several useful results on the convergence of the E.S.D. from general sample covariance matrices. In §4, we examine the strong point-wise convergence of sample eigenvalues associated to spikes. We then establish CLT for these sample eigenvalues in §5 using the methodology developed in [5]. Preliminary lemmas and their proofs are gathered in the last section.

## 2. Generalized spiked population model

In a generalized spiked population model, the population covariance matrix $T_p$ takes the form

$$T_p = \begin{pmatrix} \Sigma & 0 \\ 0 & V_p \end{pmatrix},$$

where $\Sigma$ and $V_p$ are nonnegative and nonrandom Hermitian matrices of dimension $M \times M$ and $p' \times p'$, respectively, where $p' = p - M$. The submatrix $\Sigma$ has $K$ eigenvalues $\alpha_1 > \cdots > \alpha_K > 0$ of respective multiplicity $(n_k)$, and $V_p$ has $p'$ eigenvalues $\beta_{n,1} \geq \cdots \geq \beta_{n,p'}$.

Throughout the paper, we assume that the following assumptions hold.

(a) $w_{ij}$, $i, j = 1, 2, ...$ are i.i.d. complex random variables with $Ew_{11} = 0$, $E|w_{11}|^2 = 1$, and $E|w_{11}|^4 < \infty$.

(b) $n = n(p)$ with $y_n = p'/n \to y > 0$ as $n \to \infty$.

(c) The sequence of ESD $H_n$ of $(T_p)$, i.e. generated by the population eigenvalues $\{\alpha_k, \beta_{n,j}\}$, weakly converges to a probability distribution $H$ as $n \to \infty$.

(d) The sequence $(\|T_p\|)$ of spectral norms of $(T_p)$ is bounded.

For any measure $\mu$ on $\mathbb{R}$, we denote by $\Gamma_\mu$ the support of $\mu$, a close set.

**Definition 2.1.** *An eigenvalue $\alpha$ of the matrix $\Sigma$ is called a* generalized spike eigenvalue *if $\alpha \notin \Gamma_H$.*

To avoid confusion between spikes and non-spike eigenvalues, we further assume that

(e) $\max_{1 \le j \le p'} d(\beta_{nj}, \Gamma_H) = \varepsilon_n \to 0$,

where $d(x, A)$ denotes the distance of a point $x$ to a set $A$. Note that there is a positive constant $\delta$ such that $d(\alpha_k, \Gamma_H) > \delta$, for all $k \le K$.

The above definition for generalized spikes is consistent with Johnstone's original one of (ordinary) spikes, since in that case we have $H_n \equiv H = \delta_{\{1\}}$ and $\alpha \notin \Gamma_H$ simply means $\alpha \neq 1$.

Let us decompose the observation vectors $\mathbf{x}_j = T_p^{1/2} \mathbf{u}_j$, $j = 1, \ldots, n$, where $\mathbf{u}_j = (w_{ij})_{1 \le i \le p}$ by blocs,

$$\mathbf{x}_j = \begin{pmatrix} \boldsymbol{\xi}_j \\ \boldsymbol{\eta}_j \end{pmatrix}, \quad \text{with} \quad \boldsymbol{\xi}_j = \Sigma^{1/2} (w_{ij})_{1 \le i \le M}, \quad \boldsymbol{\eta}_j = V_p^{1/2} (w_{ij})_{M < i \le p}.$$

Note that both sequences $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n\}$ and $\{\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n\}$ are i.i.d. sequences. We also denote the coordinates of $\boldsymbol{\xi}_1$ by $\boldsymbol{\xi}_1 = (\xi(1), \ldots, \xi(M))^T$.

Similarly, the sample covariance matrix $S_n = \frac{1}{n} T_p^{1/2} Z_n Z_n^* T_p^{1/2}$ is decomposed as

$$S_n = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} X_1 X_1^* & X_1 X_2^* \\ X_2 X_1^* & X_2 X_2^* \end{pmatrix},$$

with

$$X_1 = \frac{1}{\sqrt{n}} (\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_n)_{M \times n} = \frac{1}{\sqrt{n}} \boldsymbol{\xi}_{1:n}, \; X_2 = \frac{1}{\sqrt{n}} (\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_n)_{p' \times n} = \frac{1}{\sqrt{n}} \boldsymbol{\eta}_{1:n}.$$

Throughout the paper and for any Hermitian matrix $A$, we order its eigenvalues in an descending order as $\lambda_1^A \ge \lambda_2^A \ge \cdots$. By definition, the sample eigenvalues $\{\lambda_j^{S_n}, 1 \le j \le p\}$ are solutions to the equation

$$0 = |\lambda I - S_n| = |\lambda I - S_{22}| \; |\lambda I - K_n(\lambda)| \; , \tag{2.1}$$

with a random sesquilinear form

$$K_n(\lambda) = S_{11} + S_{12}(\lambda I - S_{22})^{-1} S_{21}. \tag{2.2}$$

Note that the factorization (2.1) holds for any $\lambda \notin \mathrm{spec}(S_{22})$. This identity will play a central role in our analysis.

## 3. Known results on the spectrum of large sample covariance matrices

3.1. **Marčenko-Pastur distributions.** In this section $y$ is an arbitrary positive constant and $H$ an arbitrary probability measure on $\mathbb{R}^+$. Define on the set

$$\mathbb{C}^+ := \{ z \in \mathbb{C} \; : \; \Im(z) > 0 \; \} \; ,$$

the map

$$g(s) = g_{y,H}(s) = -\frac{1}{s} + y \int \frac{t}{1+ts} dH(t) \; , \quad s \in \mathbb{C}^+. \tag{3.1}$$

It is well-known ([4, Chap. 5]) that $g$ is a one-to-one map from $\mathbb{C}^+$ onto itself, and the inverse map $m = g^{-1}$ corresponds to the Stieltjies transform of a probability measure $F_{y,H}$ on $[0, \infty)$. Throughout the paper and with a small abuse of language, we refer $F_{y,H}$ as the Marčenko-Pastur (M.P.) distribution with indexes $(y, H)$.

This family of distributions arises naturally as follows. Consider a companion matrix $\underline{S}_n = \frac{1}{n} Z_n^* T_p Z_n$ of the sample covariance matrix $S_n$. The spectra of $S_n$ and $\underline{S}_n$ are identical except $|n-p|$ zeros. It is then well-known ([11],[4, Chap. 5]) that under Conditions (a)-(d), the E.S.D. of $\underline{S}_n$ converges to the M.P. distribution $F_{y,H}$. The terminology is slightly ambiguous since the classical M.P. distribution refers to the limit of the E.S.D. of $S_n$ when $T_p = I_p$.

Note that we shall always extend a function $h$ defined on $\mathbb{C}^+$ to the real axis $\mathbb{R}$ by taking the limits $\lim_{\varepsilon \to 0_+} h(x + i\varepsilon)$ for real $x$'s whenever these limits exist. For $\alpha \notin \Gamma_H$ and $\alpha \neq 0$ define

$$\psi(\alpha) = \psi_{y,H}(\alpha) := g(-1/\alpha) = \alpha + y\alpha \int \frac{t}{\alpha - t} dH(t) \ . \tag{3.2}$$

Note that even though this formula could be extended to $\alpha = 0$ when $0 \notin \Gamma_H$, as we will see below that $\alpha$ is related to the $-1/m$ where $m$ is a Stieltjies transform, so that there is no much meaning for $\alpha = 0$. Therefore, the point 0 will always be excluded from the domain of definition of $\psi$.

Analytical properties of $F_{y,H}$ can be derived from the fundamental equation (3.2). The following lemma, due to Silverstein and Choi [14], characterizes the close relationship between the supports of the generating measure $H$ and the generated M.P. distribution $F_{y,H}$.

**Lemma 3.1.** *If $\lambda \notin \Gamma_{F_{y,H}}$, then $m(\lambda) \neq 0$ and $\alpha = -1/m(\lambda)$ satisfies*

(i) $\alpha \notin \Gamma_H$ *and* $\alpha \neq 0$ *(so that $\psi(\alpha)$ is well-defined);*

(ii) $\psi'(\alpha) > 0$.

*Conversely, if $\alpha$ satisfies* (i)-(ii)*, then $\lambda = \psi(\alpha) \notin \Gamma_{F_{y,H}}$.*

It is then possible to determine the support of $F_{y,H}$ by looking at intervals where $\psi' > 0$. As an example, Figure 1 displays the function $\psi$ for the M.P. distribution with indexes $y = 0.3$ and $H$ the uniform distribution on the set $\{1, 4, 10\}$. The function $\psi$ is strictly increasing on the following intervals: $(-\infty,\ 0)$, $(0,\ 0.63)$, $(1.40,\ 2.57)$ and $(13.19,\ \infty)$. According to Lemma 3.1, we get

$$\Gamma_{F_{y,H}} \cap \mathbb{R}^* = (0,\ 0.32) \cup (1.37,\ 1.67) \cup (18.00,\ \infty).$$

Hence, taking into account that 0 belongs to the support of $F_{y,H}$, we have

$$\Gamma_{F_{y,H}} = \{0\} \cup [0.32,\ 1.37] \cup [1.67,\ 18.00].$$

We refer to Bai and Silverstein [3] for a complete account of analytical properties of the family of M.P. distributions $\{F_{y,H}\}$ and the maps $\{\psi_{y,H}\}$. In particular, the following conclusions will be useful:

- when restricted to $\Gamma^c_{F_{y,H}}$, $\psi_{y,H}$ has a well-defined inverse function $\psi^{-1}_{y,H}$: $\Gamma^c_{F_{y,H}} \to \Gamma^c_H$ which is strictly increasing;

- the family $\{F_{y,H}\}$ is continuous in its index parameters $(y, H)$ in a wide sense. For example, $\{\psi_{y,H}\}$ tends to the identity function as $y \to 0$.

3.2. **Exact separation of sample eigenvalues.** We need first quote two results of Bai and Silverstein [2, 3] on exact separation of sample eigenvalues. Recall the ESD's $(H_n)$ of $(T_p)$, $y_n = p/n$, and let $\{F_{y_n,H_n}\}$ be the sequence of associated M.P. distributions. One should not confuse the M.P. distribution $\{F_{y_n,H_n}\}$ with the E.S.D. of $\underline{S}_n$ although both converge to the M.P. distribution $F_{y,H}$ as $n \to \infty$.

**Proposition 3.1.** *Assume hold Conditions (a)-(d) and the following*

(f) *The interval $[a, b]$ with $a > 0$ lies in an open interval $(c, d)$ outside the support of $F_{y_n,H_n}$ for all large $n$.*

*Then*

$$P(\text{ no eigenvalue of } S_n \text{ appears in } [a, b] \text{ for all large } n\ ) = 1.$$

Roughly speaking, Proposition 3.1 states that a gap in the spectra of the $F_{y_n,H_n}$'s is also a gap in the spectrum of $S_n$ for large $n$. Moreover, under Condition (f), we know by Lemma 3.1, that for large $n$,

$$\psi^{-1}_{y_n,H_n}\{[a, b]\} \subset \psi^{-1}_{y_n,H_n}\{(c, d)\} \subset \Gamma^c_{H_n}\ .$$

By continuity of $F_{y_n,H_n}$ in its indexes, it follows that we have for large $n$

$$\psi^{-1}\{[a, b]\} = \psi^{-1}_{y,H}\{[a, b]\} \subset \Gamma^c_{H_n}\ .$$

In other words, it holds almost surely and for large $n$ that, $\psi^{-1}\{[a, b]\}$ contains no eigenvalue of $T_p$. Let for these $n$, the integer $i_n \geq 0$ be such that

$$T_p \text{ has exactly } i_n \text{ eigenvalues larger than } \psi^{-1}(b)\ . \tag{3.3}$$

**Proposition 3.2.** *Assume Conditions (a)-(d) and (f) hold. If $y[1-H(0)] \leq 1$, or $y[1-H(0)] > 1$ but $[a, b]$ is not contained in $[0, x_0]$ where $x_0 > 0$ is the* *smallest value of the support of* $F_{y,H}$*, then with* $i_n$ *defined in (3.3) we have*

$$P(\lambda_{i_n+1}^{S_n} \leq a < b \leq \lambda_{i_n}^{S_n} \quad \text{for all large } n) = 1.$$

In other words, under these conditions, it happens eventually that the numbers of sample eigenvalues $\{\lambda_i^{S_n}\}$ in both sides of $[a, b]$ match exactly the numbers of populations eigenvalues $\{\alpha_k, \beta_{n,j}\}$ in both sides of the interval $\psi^{-1}\{[a, b]\}$.

## 4. Almost sure convergence of sample eigenvalues from generalized spikes

From (3.2), we have

$$\psi'(\alpha) = 1 - y \int \frac{t^2}{(\alpha - t)^2} dH(t) \ , \qquad \psi'''(\alpha) = -6y \int \frac{t^2}{(\alpha - t)^4} dH(t) \ .$$

Therefore, when $\alpha$ approaches the boundary of the support of $H$, $\psi'(\alpha)$ tends to $-\infty$, see also Figure 1. Moreover, $\psi'$ is concave on any interval outside $\Gamma_H$.

As we will see, the asymptotic behavior of the sample eigenvalues generated by a generalized spike eigenvalue $\alpha$ depends on the sign of $\psi'(\alpha)$.

**Definition 4.1.** *We call a generalized spike eigenvalue* $\alpha$*, a* distant spike *for the M.P. law* $F_{y,H}$ *if* $\psi'(\alpha) > 0$*, and a* close spike *if* $\psi'(\alpha) \leq 0$*.*

Recall that $\psi$ depend on the parameters $(y, H)$. When $H$ is fixed, and since $\psi$ tends to the identity function as $y \to 0$, a close spike for a given M.P. law $F_{y,H}$ becomes a distant spike for M.P. law $F_{y,H}$ for small enough $y$.

As an example, different types of spikes are displayed in Figure 2. The solid curve corresponds to a zoomed view of $\psi_{0.3,H}$ of Figure 1. For $F_{0.3,H}$, the three values $\alpha_1$, $\alpha_2$ and $\alpha_5$ are close spikes; each small enough $\alpha$ (close to zero), or large enough $\alpha$ (not displayed), or a value between $u$ and $v$ (see the figure) is a distant spike. Furthermore, as $y$ decreases from 0.3 to 0.02 (dashed curve), $\alpha_1$, $\alpha_2$ and $\alpha_5$ become all distant spikes.

Throughout this section, for each spike eigenvalue $\alpha_k$, we denote by $\nu_k+1,\ldots,\nu_k+n_k$ the descending ranks of $\alpha_k$ among the eigenvalues of $T_p$ (multiplicities of eigenvalues are counted): in other words, there are $\nu_k$ eigenvalues of $T_p$ larger than $\alpha_k$ and $p-\nu_k-n_k$ less.

**Theorem 4.1.** *Assume that the conditions (a)-(e) hold. Let $\alpha_k$ be a generalized spike eigenvalue of multiplicity $n_k$ satisfying $\psi'(\alpha_k)>0$ (distant spike) with descending ranks $\nu_k+1,\cdots,\nu_k+n_k$. Then, the $n_k$ consecutive sample eigenvalues $\{\lambda_i^{S_n}\}$, $i=\nu_k+1,\ldots,\nu_k+n_k$ converge almost surely to $\psi(\alpha_k)$.*

*Proof.* Recall Figure 2 of the $\psi$ function, for each distant spike $\alpha_k$, there is an interval $(u_k,v_k)$ such that

- $u_k<\alpha_k<v_k$;
- $\psi'(u_k)=\psi'(v_k)=0$;
- $\psi'(\alpha)>0$ for all $\alpha\in(u_k,v_k)$.

Here we make the convention that $v_k=\infty$ if $\psi'(\alpha)>0$ for all $\alpha>\alpha_k$ and $u_k=0$ if $\psi'(\alpha)>0$ for all $\alpha\in(0,\alpha_k)$.

Recall that the support of $F_{y_n,H_n}$ is determined by

(4.1)
$$\psi_n'(\alpha)=\psi'_{y_n,H_n}(\alpha)=1-y_n\left[\frac{p'}{p}\int\frac{t^2}{(\alpha-t)^2}dH_n^v(t)\ +\frac{1}{p}\sum_{j=1}^K\frac{n_j\alpha_j^2}{(\alpha-\alpha_j)^2}\right],$$

where $H_n^v=\frac{1}{p'}\sum_j\delta_{\beta_{n,j}}$ is the ESD of $V_p$.

Let $\tilde v_k=\min(v_k,\alpha_{k-1})$ if $k>1$ and $\tilde v_k=v_k$ otherwise. Choose $v,v'$ and $\alpha_u',\alpha_u$ such that $\alpha_k<\alpha_u'<\alpha_u<v<v'<\tilde v_k$. By condition (e), all eigenvalues of $T_p$ will keep away from the interval $(\alpha_u',v')$ for all large $n$. Thus, $\psi_n'(\alpha)\to\psi'(\alpha)>0$ uniformly on the interval $[\alpha_u',v']$. Hence, the interval $(\psi(\alpha_u'),\psi(v'))$ will be out of the support of $F_{y_n,H_n}$ for all large $n$. Consequently, the interval $[\psi(\alpha_u),\psi(v)]$ satisfies the conditions of Proposition 3.2 with $i_n=\nu_k$. Therefore, by Proposition 3.2, we have

$$\begin{cases}P(\lambda_{\nu_k+1}^{S_n}\le\psi(\alpha_u)<\psi(v)\le\lambda_{\nu_k}^{S_n},\ \text{ for all large } n)=1 & \text{if } \nu_k>0;\\ P(\lambda_{\nu_k+1}^{S_n}\le\psi(\alpha_u),\ \text{ for all large } n)=1 & \text{otherwise.}\end{cases}$$

Therefore, it holds almost surely

$$\limsup_n \lambda^{S_n}_{\nu_k+1} \le \psi(\alpha_u),$$

and finally, letting $\alpha_u \to \alpha_k$,

$$\limsup_n \lambda^{S_n}_{\nu_k+1} \le \psi(\alpha_k). \tag{4.2}$$

Similarly, one can prove that for any $\tilde{u}_k < u < \alpha_l < \alpha_k$,

$$\begin{cases} P(\lambda^{S_n}_{\nu_k+n_k+1} \le \psi(u) < \psi(\alpha_l) \le \lambda^{S_n}_{\nu_k+n_k}, \ \text{ for all large } n) = 1 & \text{if } \nu_k+n_k<p, \\ P(\lambda^{S_n}_{\nu_k+n_k} \ge \psi(\alpha_l), \ \text{ for all large } n) = 1 & \text{otherwise,} \end{cases}$$

where $\tilde{u}_k = \max(u_k, \alpha_{k+1})$ if $k < K$ and $\tilde{u}_k = u_k$ otherwise.

Consequently,

$$\liminf_n \lambda^{S_n}_{\nu_k+n_k} \ge \psi(\alpha_k). \tag{4.3}$$

Thus, we proved that almost surely,

$$\lim_n \lambda^{S_n}_{\nu_k+j} = \psi(\alpha_k), \ \text{ for } j = 1, \cdots, n_k.$$

The proof of Theorem 4.1 is complete. □

Next we consider close spikes.

**Theorem 4.2.** *Assume that the conditions (a)-(e) hold. Let $\alpha_k$ be a generalized spike eigenvalue of multiplicity $n_k$ satisfying $\psi'(\alpha_k) \le 0$ (close spike) with descending ranks $\nu_k+1, \ldots, \nu_k+n_k$. Let $I$ be the maximal interval in $\Gamma^c_H$ containing $\alpha_k$.*

(i) *If $I$ has a sub-interval $(u_k, v_k)$ on which $\psi' > 0$ (then we take this interval to be maximal), then the $n_k$ sample eigenvalues $\{\lambda^{S_n}_j\}$, $j = \nu_k+1, \ldots, \nu_k+n_k$ converge almost surely to the number $\psi(w)$ where $w$ is one of the endpoints $\{u_k, v_k\}$ nearest to $\alpha_k$ ;*

(ii) *If for all $\alpha \in I$, $\psi'(\alpha) \le 0$, then the $n_k$ sample eigenvalues $\{\lambda^{S_n}_j\}$, $j = \nu_k+1, \ldots, \nu_k+n_k$ converge almost surely to the $\gamma$-th quantile of $G$, the L.S.D. of $S_n$, where $\gamma = H(0, \alpha_k)$.*

*Proof.* The proof refers to the curves of Figure 2.

(i). Suppose $\alpha_k$ is a spike eigenvalue satisfying $\psi'(\alpha_k) \leq 0$ and there is an interval $(u_k, v_k) \subset I$ on which $\psi' > 0$ ($\alpha_k$ is like the $\alpha_1$ on the figure). According to Lemma 3.1, $\psi\{(u_k, v_k)\} \subset \Gamma^c_{F_{y,H}}$ and $\psi(u_k)$ is a boundary point of the support of $G$, the L.S.D. of $S_n$. Without loss of generality, we can assume $\alpha_k \leq u_k$, the argument of the other situation where $\alpha_k > v_k$ being similar.

Choose $u_k < \alpha_u < v < \tilde{v}$ ($\tilde{v} = \min(v_k, \alpha_{k-1})$ or $v_k$ in accordance with $k > 1$ or not) such that $(\alpha_u, v) \subset I$, by the argument used in the proof of Theorem 4.1, one can prove that

$$\begin{cases} P(\lambda^{S_n}_{\nu_k+1} \leq \psi(\alpha_u) < \psi(v) \leq \lambda^{S_n}_{\nu_k}, \quad \text{for all large } n) = 1 & \text{if } \nu_k > 0; \\ P(\lambda^{S_n}_{\nu_k+1} \leq \psi(\alpha_u), \quad \text{for all large } n) = 1 & \text{otherwise.} \end{cases}$$

This proves that almost surely,

$$\limsup \lambda^{S_n}_{\nu_k+1} \leq \psi(u_k) \leq \liminf \lambda^{S_n}_{\nu_k} \ .$$

On the other hand, since $\psi(u_k)$ is a boundary point of the support of $G$, we know that for any $\varepsilon > 0$, almost surely, the number of $\lambda^{S_n}_i$'s falling into $[\psi(u_k) - \varepsilon, \psi(u_k)]$ tends to infinity. Therefore,

$$\liminf \lambda^{S_n}_{\nu_k+n_k+1} \geq \psi(u_k) - \varepsilon, \quad \text{a.s.}.$$

Since $\varepsilon$ is arbitrary, we have finally proved that almost surely,

$$\lim \lambda^{S_n}_{\nu_k+j} = \psi(u_k), \quad j = 1, \cdots, n_k.$$

Thus, the proof of Conclusion (i) of Theorem 4.2 is complete.

Similarly, if the spiked eigenvalue $\alpha_k$ is like $\alpha_2$, we can show that the $n_k$ corresponding eigenvalues of $S_n$ goes to $\psi(v_k)$.

(ii) If the spiked eigenvalues is like $\alpha_5$, where the gap of support of LSD disappeared, clearly the corresponding sample eigenvalues $\lambda_{\nu_k+1}, \ldots, \lambda_{\nu_k+n_k}$ tend to the $\gamma$-th quantile of the LSD of $S_n$ where

$$\gamma = 1 - \lim \frac{i_n}{\nu_k} = H(0, \alpha_k).$$

□

4.1. **Case of Johnstone's spiked population model.** In the case of Johnstone's model, $H$ reduces to the Dirac mass $\delta_1$ and the LSD $G$ equals the Marčenko-Pastur law with $\Gamma_G = [a_y, b_y]$. Each $\alpha > 0$, $\alpha \neq 1$ is then a spike eigenvalue. The associated function $\psi$ in (3.2) becomes

$$\psi(\alpha_k) = \alpha_k + \frac{y\alpha_k}{\alpha_k - 1}. \tag{4.4}$$

The function $\psi$ has the following properties, see Figure 3:

- its range equals $(-\infty, a_y] \cup [b_y, \infty)$ ;
- $\psi(1-\sqrt{y}) = a_y$ , $\psi(1+\sqrt{y}) = b_y$;
- $\psi'(\alpha) > 0 \Leftrightarrow |\alpha - 1| > \sqrt{y}$.

Therefore, by Theorem 4.1, for any spike eigenvalue satisfying $\alpha_k > 1+\sqrt{y}$ (large enough) or $\alpha_k < 1-\sqrt{y}$ (small enough), there is a packet of $n_k$ consecutive eigenvalues $\{\lambda_{n,j}\}$ converging almost surely to $\psi(\alpha_k) \notin [a_y, b_y]$. In other words, assume there are exactly $K_1$ spikes greater than $1+\sqrt{y}$ and $K_2$ spikes smaller than $1-\sqrt{y}$. By Theorems 4.1 and 4.2 we conclude that

(i) the $N_1 := n_1 + \ldots + n_{K_1}$ largest eigenvalues $\{\lambda_j^{S_n}\}$, $j = 1, \ldots, N_1$ tend to their respective limits $\{\psi(\alpha_k)\}$, $k = 1, \ldots, K_1$ ;
(ii) the immediately following largest eigenvalue $\lambda_{N_1+1}^{S_n}$ tends to the right edge $b_y$;
(iii) the $N_2 := n_K + \cdots + n_{K-K_2+1}$ smallest sample eigenvalues $\{\lambda_{n,p-j}^{S_n}\}$, $j = 0, \ldots, N_2 - 1$ tend to their respective limits $\{\psi(\alpha_k)\}$, $k = K, \ldots, K - K_2 + 1$ ;
(iv) the immediately following smallest eigenvalue $\lambda_{p-N_2}^{S_n}$ tends to the left edge $a_y$.

Hence we have recovered the content of Theorem 1.1 of [6].

4.2. **An example of generalized spike eigenvalues.** Assume that $T_p$ is diagonal with three base eigenvalues $\{1, 4, 10\}$, nearly $p/3$ times for each of them, and there are four spike eigenvalues $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (15,\ 6,\ 2,\ 0.5)$, with respective multiplicities $(n_k) = (3, 2, 2, 2)$. The limiting population-sample ratio is taken to be $y = 0.3$. The limiting population spectrum $H$

is then the uniform distribution on $\{1, 4, 10\}$. The support of the limiting Marčenko-Pastur distribution $F_{0.3,H}$ contains two intervals [0.32, 1.37] and [1.67, 18], see §3.1. The $\psi$-function of (3.2) for the current case is displayed in Figure 1. For simulation, we use $p' = 600$ so that $T_p$ has the following 609 eigenvalues:

$$15,\ 15,\ 15,\ \underbrace{10, \ldots, 10}_{200},\ 6,\ 6,\ \underbrace{4, \ldots, 4}_{200},\ 2,\ 2, \underbrace{1, \ldots, 1}_{200},\ 0.5,\ 0.5\ .$$

From the table

| spike $\alpha_k$ | 15 | 6 | 2 | 0.5 |
|---|---|---|---|---|
| multiplicity $n_k$ | 3 | 2 | 2 | 2 |
| $\psi'(\alpha_k)$ | $+$ | $-$ | $+$ | $-$ |
| $\psi(\alpha_k)$ | 18.65 | 5.82 | 1.55 | 0.29 |
| descending ranks | 1, 2, 3 | 204, 205 | 406, 407 | 608, 609 |

we see that 6 is a close spike for $H$ while the three others are distant ones. By Theorems 4.1 and 4.2, we know that

- the 7 sample eigenvalues $\lambda_j^{S_n}$ with $j \in \{1,\ 2,\ 3,\ 406,\ 407,\ 608,\ 609\}$ associated to distant spikes tend to 18.65, 1.55 and 0.29, respectively, which are located outside the support of limiting distribution $F_{0.3,H}$ (or $G$);
- the two sample eigenvalues $\lambda_j^{S_n}$ with $j = 204, 205$ associated to the close spike 6 tend to a limit located inside the support, the $\gamma$-th quantile of the limiting distribution $G$ where $\gamma = H(0, 6) = 2/3$.

There facts are illustrated by a simulation sample displayed in Figure 4.

## 5. CLT for sample eigenvalues from distant generalized spikes

Following Theorem 4.1, to any distant generalized spike eigenvalue $\alpha_k$, there is a packet of $n_k$ consecutive sample eigenvalues $\{\lambda_j^{S_n}\ :\ \ j \in J_k\}$ converging to $\psi(\alpha_k) \notin \Gamma_G$ where $J_k$ are the descending ranks of $\alpha_k$ among the eigenvalues of $T_p$ (counting multiplicities). The aim of this section is to derive a CLT for $n_k$-dimensional vector

$$\sqrt{n}\{\lambda_j^{S_n} - \psi(\alpha_k)\}\ , \quad j \in J_k.$$

The method follows Bai and Yao [5] which considers Johnstone's spiked population model. Consider the random form $K_n$ introduced in (2.2) and let

$$A_n = (a_{ij}) = A_n(\lambda) = X_2^*(\lambda I - X_2 X_2^*)^{-1} X_2, \quad \lambda \notin \Gamma_G. \tag{5.1}$$

By Lemma 6.2, detailed in §6, we know that $n^{-1} tr A_n$, $n^{-1} tr A_n A_n^*$ and $n^{-1} \sum_{i=1}^n a_{ii}^2$ converge, almost surely or in probability, to $ym_1(\lambda)$, $ym_2(\lambda)$ and $(y[1+m_1(\lambda)]/\{\lambda - y[1+m_1(\lambda)]\})^2$, respectively. Here, the $m_j(\lambda)$ are some specific transforms of the LSD $G$ (see §6).

Therefore, the random form $K_n$ in (2.2) can be decomposed as follows

$$\begin{aligned} K_n(\lambda) &= S_{11} + X_1 A_n X_1^* = \frac{1}{n} \xi_{1:n}(I + A_n)\xi_{1:n}^* \\ &= \frac{1}{n} \left\{ \xi_{1:n}(I + A_n)\xi_{1:n}^* - \Sigma \mathrm{tr}(I + A_n) \right\} + \frac{1}{n} \Sigma \mathrm{tr}(I + A_n) \\ &= \frac{1}{\sqrt{n}} R_n + [1 + ym_1(\lambda)] \, \Sigma + o_P(\frac{1}{\sqrt{n}}), \end{aligned}$$

with

$$R_n = R_n(\lambda) = \frac{1}{\sqrt{n}} \left\{ \xi_{1:n}(I + A_n)\xi_{1:n}^* - \Sigma \mathrm{tr}(I + A_n) \right\}. \tag{5.2}$$

In the last derivation, we have used the fact

$$\frac{1}{n} tr(I + A_n) = 1 + ym_1(\lambda) + o_P(\frac{1}{\sqrt{n}}),$$

which follows from a CLT for $\mathrm{tr}(A_n)$ [see 1].

For the statement of our result, we first need to find the limit distribution of the sequence of random matrices $\{R_n(\lambda)\}$. The situation is different for the real and complex cases. By applications of Propositions 3.1 and 3.2 in [5], we have for $\lambda \notin \Gamma_G$,

(i) if the variables $(w_{ij})$ are real-valued, the random matrix $R_n(\lambda)$ converges weakly to a symmetric random matrix $R(\lambda) = (R_{ij}(\lambda))$ with zero-mean Gaussian entries having an explicitly known covariance function ;

(ii) if the variables $(w_{ij})$ are complex-valued, the random matrix $R_n$ converges weakly to a zero-mean Hermitian random matrix $R(\lambda) =$

$(R_{ij}(\lambda))$. Moreover, the real and imaginary parts of its upper-triangular bloc $\{R_{ij}(\lambda),\ 1 \le i \le j \le M\}$ form a $2K$-dimensional Gaussian vector with an explicitly known covariance matrix.

We are in order to introduce our CLT. Let the spectral decomposition of $\Sigma$,

$$\Sigma = U \begin{pmatrix} \alpha_1 I_{n_1} & \cdots & 0 \\ 0 & \ddots & 0 \\ \cdots & 0 & \alpha_K I_{n_K} \end{pmatrix} U^* \ , \tag{5.3}$$

where $U$ is an unitary matrix. Let $\psi_k = \psi(\alpha_k)$ and $R(\psi_k)$ be the weak Gaussian limit of the sequence of matrices of random forms $[R_n(\psi_k)]_n$ recalled above (in both real and complex variables case). Let

$$\widetilde{R}(\psi_k) = U^* R(\psi_k) U \ . \tag{5.4}$$

**Theorem 5.3.** *For each distant generalize spike eigenvalue, the $n_k$-dimensional real vector*

$$\sqrt{n}\{\lambda_j^{S_n} - \psi_k,\ j \in J_k\} \ ,$$

*converges weakly to the distribution of the $n_k$ eigenvalues of the Gaussian random matrix*

$$\frac{1}{1 + y m_3(\psi_k)\alpha_k} \widetilde{R}_{kk}(\psi_k).$$

*where $\widetilde{R}_{kk}(\psi_k)$ is the $k$-th diagonal block of $\widetilde{R}(\psi_k)$ corresponding to the indexes $\{u, v \in J_k\}$.*

It is worth noticing that the limiting distribution of such $n_k$ packed sample extreme eigenvalues are generally *non Gaussian* and asymptotically dependent. Indeed, the limiting distribution of a single sample extreme eigenvalue $\lambda_j^{S_n}$ is Gaussian if and only if the corresponding generalized spike eigenvalue is simple. We refer the reader to [5] for detailed examples illustrating these same facts but for Johnstone's model.

## 6. Lemmas

For $\lambda \notin \Gamma_G$, we define

$$\begin{aligned} m_1(\lambda) &= \int \frac{x}{\lambda - x} dG(x), \\ m_2(\lambda) &= \int \frac{x^2}{(\lambda - x)^2} dG(x) \ , \\ m_3(\lambda) &= \int \frac{x}{(\lambda - x)^2} dG(x) \ . \end{aligned}$$

The following lemma gives the law of large numbers for some useful statistics of $A_n$ defined in (5.1). We omit its proof because it is a straightforward extension of Lemma 6.1 of [5], related to Johnstone's spiked population model, to the present generalized spiked population model.

**Lemma 6.2.** *Under the assumptions of Theorem 4.1, for all* $\lambda \in [a, b]$*, we have*

$$\frac{1}{n} trA_n \xrightarrow{a.s.} ym_1(\lambda) \ , \tag{6.1}$$

$$\frac{1}{n} trA_n A_n^* \xrightarrow{a.s.} ym_2(\lambda) \ , \tag{6.2}$$

$$\frac{1}{n} \sum_{i=1}^{n} a_{ii}^2 \xrightarrow{a.s.} \left( \frac{y[1 + m_1(\lambda)]}{\lambda - y[1 + m_1(\lambda)]} \right)^2 . \tag{6.3}$$

**Lemma 6.3.** *For all* $\lambda \in [a, b]$, $K_n(\lambda)$ *converges almost surely to the constant matrix* $[1 + ym_1(\lambda)]\Sigma$.

*Proof.* The random form $K_n$ in (2.2) can be decomposed as follows

$$K_n(\lambda) = S_{11} + X_1 A_n X_1^* = \frac{1}{n} (\xi_1, \dots, \xi_n)(I + A_n)(\xi_1, \dots, \xi_n)^* .$$

Define $M$ be the event that $S_{22}$ has no eigenvalues in the interval $[a', b']$ which satisfies $[a, b] \subset (a', b')$ and $[a', b'] \subset (c, d)$. On the event $M$, the norm of $A_n$ is bounded by $\max\{\frac{1}{a - a'}, \frac{1}{b' - b}\}$. By independence, it is easy to show that

$$\frac{1}{n} \left\{ (u_1, \dots, u_n)(I + A_n)(u_1, \dots, u_n)^* I_M - [\mathrm{tr}(I + A_n)] I_M \right\} \overset{a.s.}{\to} 0.$$

By proposition 3.1, $I_m \to 1, a.s.$. Thus

$$D_n(\lambda) = o_{a.s.}(1) + + [\frac{1}{n} \mathrm{tr}(I + A_n)] \Sigma I_M \overset{a.s.}{\to} (1 + ym_1(\lambda))\Sigma, \tag{6.4}$$

where the last step follows from (6.1). The conclusion follows. □

## References


[1] Z.D. Bai and J.W. Silverstein. CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.*, 32:553–605, 2004.

[2] Z.D. Bai and J.W. Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large dimensional sample covariance matrices. *Ann. Probab.*, 26:316–345, 1998.

[3] Z.D. Bai and J.W. Silverstein. Exact separation of eigenvalues of large dimensional sample covariance matrices. *Ann. Probab.*, 27(3):1536–1555, 1999.

[4] Z.D. Bai and J.W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices.* Science Press, Beijing, 2006.

[5] Z.D. Bai and J.-F. Yao. Central limit theorems for eigenvalues in a spiked population model. *Ann. Inst. Henri Poincaré*, 44(3):447–474, 2008.

[6] J. Baik and J.W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate. Anal.*, 97:1382–1408, 2006.

[7] J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, 33(5):1643–1697, 2005.

[8] M. Capitaine, C. Donati-Martin, and D. Féral. The largest eigenvalue of finite rank deformation of large wigner matrices: convergence and non-universality of the fluctuations. Technical report, `arXiv:math/0605624`, 2007.

[9] D. Féral and S. Péché. The largest eigenvalue of rank one deformation of large Wigner matrices. *Comm. Math. Phys.*, 272(1):185–228, 2007.

[10] I. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statistics*, 29(2):295–327, 2001.

[11] V.A. Marčenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sb*, 1:457–483, 1967.

[12] S. Péché. The largest eigenvalue of small rank perturbations of Hermitian random matrices. *Probab. Theory Related Fields*, 134(1):127–173, 2006.

[13] Jack W. Silverstein. Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.*, 55(2):331–339, 1995.

[14] Jack W. Silverstein and Sang-Il Choi. Analysis of the limiting spectral distribution of large-dimensional random matrices. *J. Multivariate Anal.*, 54(2):295–309, 1995.

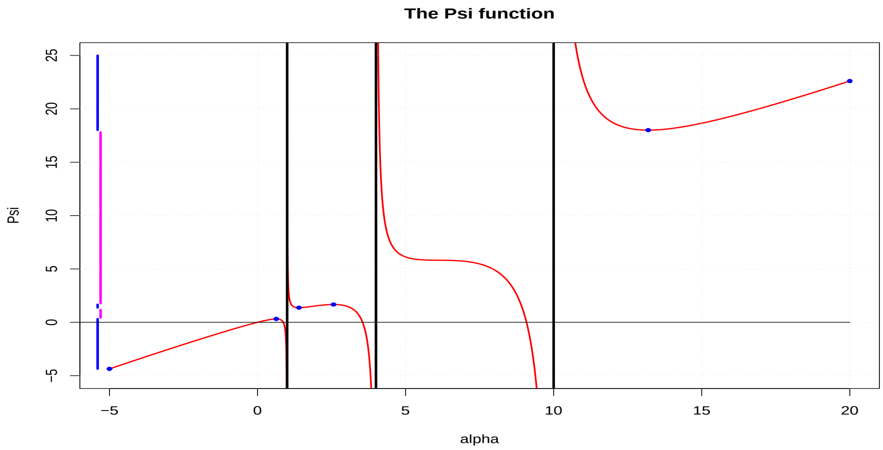


FIGURE 1. The $\psi$ function for the Marčenko-Pastur distribution $F_{0.3,H}$ with $H$ the uniform distribution on the set $\{1, 4, 10\}$. Blue points indicate intervals where $\psi' > 0$. Singular points of $\psi$ are indicated as vertical lines corresponding to the support of $H$. On the left, the support set of $F_{0.3,H}$ (except the point 0) and its complementary set are indicated as magenta and blue segments respectively.

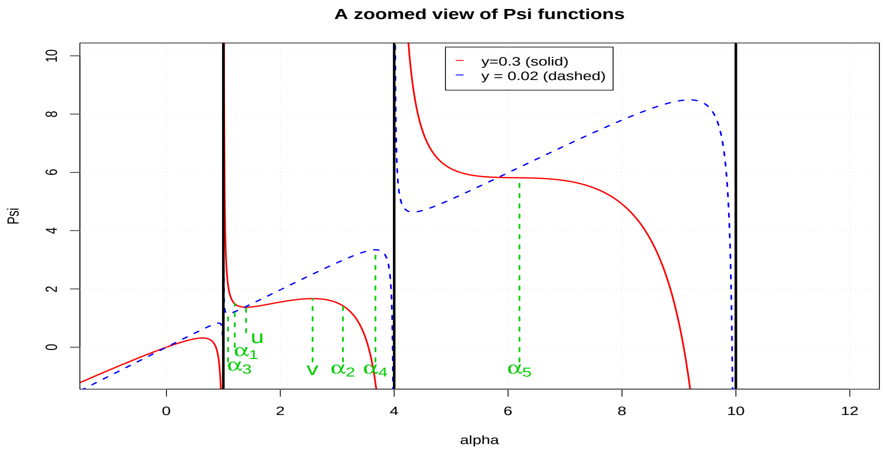


FIGURE 2. A zoomed view of the $\psi$ functions for the Marčenko-Pastur distribution $F_{0.3,H}$ (solid curve) and $F_{0.02,H}$ (dashed curve) with $H$ the uniform distribution on the set $\{1, 4, 10\}$. The three points $\alpha_1$, $\alpha_2$ and $\alpha_5$ are close spikes for $F_{0.3,H}$ where $\psi'_{0.3,H} \leq 0$. They become all distant spikes for $F_{0.02,H}$ as $\psi'_{0.02,H} > 0$.

ZHIDONG BAI, KLASMOE AND SCHOOL OF MATHEMATICS AND STATISTICS, NORTHEAST NORMAL UNIVERSITY, 5268 PEOPLE'S ROAD, 130024 CHANGCHUN, CHINA, AND, DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY, NATIONAL UNIVERSITY OF SINGAPORE, 10, KENT RIDGE CRESCENT, SINGAPORE 119260

*E-mail address*: `stabaizd@nus.edu.sg`

JIAN-FENG YAO, IRMAR AND UNIVERSITÉ DE RENNES 1, CAMPUS DE BEAULIEU, 35042 RENNES CEDEX, FRANCE,

*E-mail address*: `jian-feng.yao@univ-rennes1.fr`

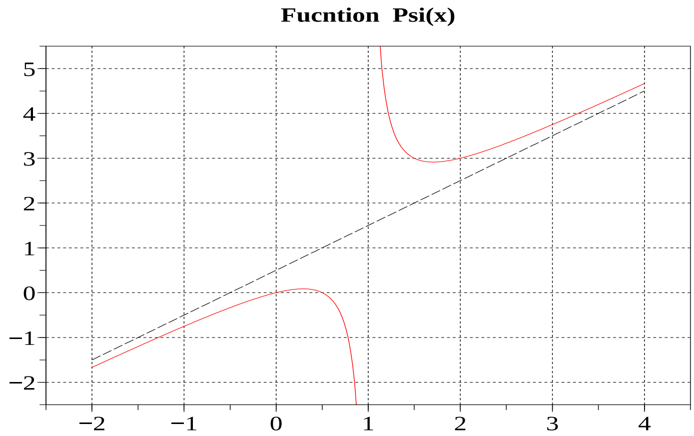


FIGURE 3. The function $\alpha \mapsto \psi(\alpha) = \alpha + y\alpha/(\alpha - 1)$ which maps a spike eigenvalue $\alpha$ to the limit of an associated sample eigenvalue in Johnstone's spiked population model. Figure with $y = \frac{1}{2}$; $[1 \mp \sqrt{y}] = [0.293, \ 1.707]$; $[(1 \mp \sqrt{y})^2] = [0.086, \ 2.914]$ .

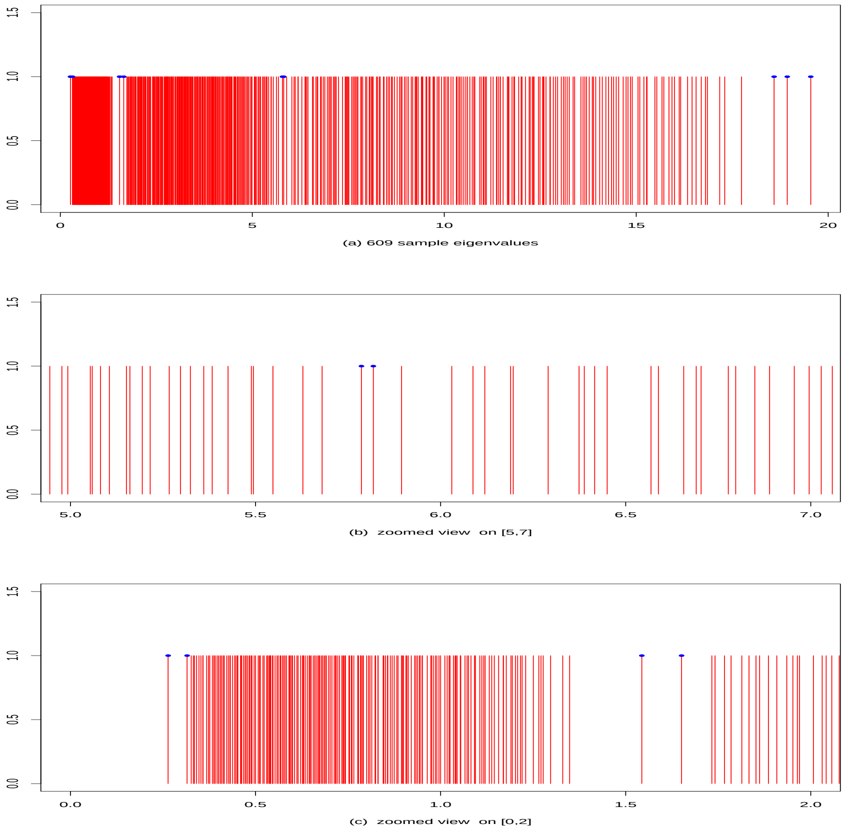


FIGURE 4. An example of $p = 609$ sample eigenvalues (a), and two zoomed views (b) and (c) on [5,7] and [0,2] respectively. The limiting distribution of the E.S.D has support $[0.32, 1.37] \cup [1.67, 18.00]$. The 9 sample eigenvalues $\{\lambda_j^{S_n},\ j = 1, 2, 3, 204, 205, 406, 407, 608, 609\ \}$ associated to the spikes are marked with a blue point. Gaussian entries.